DRAWING SENSITIVITY CURVES: ASK A SILLY QUESTION, GET A SILLY ANSWER

R. G. Nolty
California Institute of Technology

Abstract
Sensitivity plots are meant to answer the question, "If the experiment were to see what is expected, what kind of confidence limits could the experiment be expected to set?" In practice this may be done naively by assuming the experiment will produce the data vector **V** which is the most likely data vector predicted for parameters **X** (for example, the parameters may be oscillation parameters, $\Delta m^2, sin^2(2\theta)$ and the data vector may be the number of events in angular bins) and drawing the C.L. curve in parameter space for experimental result **V**. It is often the case that some other parameters $\tilde{X}$ (that are "close" to **X** in parameter space) will predict a most likely data vector $\tilde{V}$ that is extremely similar to **V**. In that case, if the experiment produced data vector **V**, the theory $\tilde{X}$ would be excluded only at a very small C.L. (e.g., perhaps 5%). However, in real experiments with limited statistics, if the true parameters of nature were **X**, the experiment would measure a data vector that is fluctuated from **V**, and the theory $\tilde{X}$ would probably be excluded at a much higher C.L. (typically greater than 50%). Thus, the naive method does not do a good job of answering the essential question. I will present a simple algorithm that takes fluctuations into account when drawing sensitivity curves, and illustrate it with an example from atmospheric neutrino oscillations.

As far as I know, the concept of sensitivity was first formally defined in the 1998 paper of Feldman and Cousins [1], where it was defined as "the average upper limit that would be obtained by an ensemble of experiments with the expected background and no true signal". Informally, however, the definition is broader and includes sensitivity to non-null hypotheses. Perhaps a good statement of the broader informal usage of the word sensitivity is "the limit that an experiment would set if it saw what was expected", where the expectation may refer to a null or a non-null hypothesis. For example, a proposal for a new neutrino oscillations experiment may include a plot labeled "sensitivity" which shows the limit that would be set if the experiment saw the data predicted by the Super Kamiokande best fit parameters. There is also at least one case in which a paper presenting an experimental result [2] included a graph labeled "sensitivity", which showed the limit that would have been expected *a priori* assuming the true parameters of nature were the best fit parameters which had been obtained by that very experiment. (In that case, the sensitivity was shown because the data vector that was actually obtained had a fairly poor fit to all hypotheses, including the best fit. The 90% C.L. curve was thus much more restrictive than the *a priori* expectation, and the experimenters felt it was only honest to point this out on the results plot.)

The point of this paper is to point out a simple trap the experimentalist may fall into when computing so-called sensitivity curves for complicated experiments. Let us consider an atmospheric neutrino oscillations experiment in which ν_μ events are accumulated in a number of angular bins. Using a model for neutrino flux and cross sections, and for a particular oscillations hypothesis (i.e. a particular choice of oscillations parameters), the experimentalist may predict the number of events that will show up in each bin. Then she may say to herself, "Let me assume that the experiment measures exactly the predicted number in each bin. What confidence limits curve would I then draw in parameter space?"

Figure 1 shows a possible result. The figure is based on a rough approximation to the MACRO experiment, with a fairly short running period of just a couple of years. The "data" from which the graph

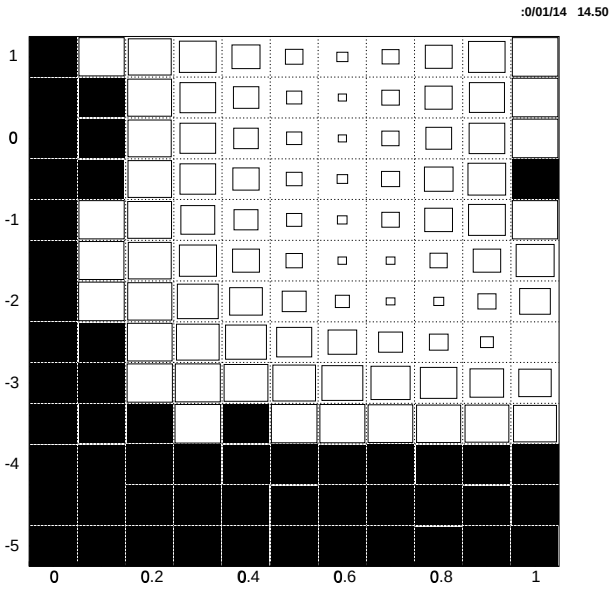

Fig. 1: Plot of exclusion levels using naive algorithm. The plot consists of a grid of points in parameter space. The x-axis is $sin^2(2\theta)$ with intervals of 0.1. The y-axis is $log_{10}(\Delta m^2)$ with intervals of 0.5. At each grid point, the size of the box is proportional to the confidence level at which that hypothesis is excluded (to be more precise, the lowest confidence level at which that hypothesis is not excluded). All of the black squares are greater than 90%, so all of those hypotheses are excluded at 90% C.L. This graph was computed for an experiment which measured exactly the prediction for $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$

was produced was the exact prediction for $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$. The figure consists of a grid of test-point hypotheses. At each grid point, the size of the square is proportional to the confidence level at which that hypothesis is excluded. The black squares are larger than 90%, and thus those hypotheses are excluded at the 90% C.L.

While the 90% C.L. exclusion curve that can be estimated from this figure appears reasonable, I wish to draw your attention to the extremely small squares running along a curve from $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$ to $\Delta m^2 = 10^{1}; \sin^2(2\theta) = 0.6$. For example, the hypothesis $\Delta m^2 = 10^{1}$; $\sin^2(2\theta) = 0.6$ is excluded only at the 5% C.L. This does not seem plausible for a real experiment.

The source of the problem is made apparent in Figure 2. In this figure, the data bins (the prediction for $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$) are given by the thick line, while the thin line with error bars gives the prediction for a test hyptothesis $\Delta m^2 = 10^{1}; \sin^2(2\theta) = 0.6$. While a large statistics experiment could distinguish between the two due to their different slopes, at the current statistics, a simple chi-squared evaluation would show the "data" agreeing with the test hypothesis much better than a dataset generated with fluctuations from the test hypothesis could be expected to. To be quantitative, only about 5% of datasets generated from the test hypothesis with fluctuations score better than the "data" generated without fluctuations from the default hypothesis $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$.

However, it is not likely that the real experiment would produce data in such close agreement with both predictions. If the default hypothesis were true, the experiment would produce data fluctuated from the default prediction. This data will probably agree with the test hypothesis much less than the unfluctuated data does, and thus the test hypothesis will probably be excluded at a greater C.L. than the sensitivity calculation shows. Thus, our naive algorithm does a poor job of answering the informal question, "What limit would the experiment set if it saw what was expected?"

Our naive experimenter has missed a point which Feldman and Cousins got right – a sensitivity calculation should be based on an ensemble of (fluctuated) experiments. I propose an extended definition

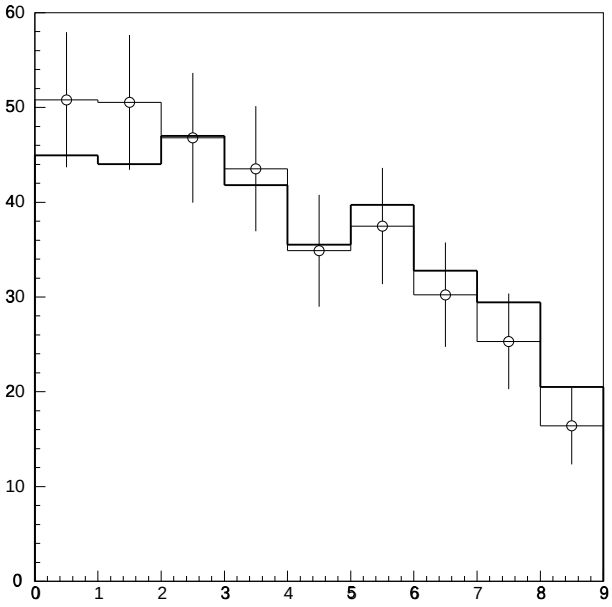

Fig. 2: The predicted bins for two hypotheses. The x-axis is bin-number for a histogram in $\cos(\theta_{zenith})$, with vertical to the left. The thick line is the prediction for the default hypothesis $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$. This was taken as the data vector measured by the experiment. The thin line with error bars is the prediction for a test point hypothesis $\Delta m^2 = 10^{1}; \sin^2(2\theta) = 0.6$. The error bars shown are $\sqrt{n}$ only.

of sensitivity, quantitatively slightly different from that of Feldman and Cousins but similar in spirit and easier to calculate in a multi-dimensional parameter space. "The sensitivity of an experiment to a default hypothesis is given by computing, at all test point hypotheses in parameter space, the average (over an ensemble of experiments fluctuated from the default hypothesis predictions) of the lowest confidence level at which the test point is not excluded."

The application of this prescription results in the exclusion plot in Figure 3. Qualitatively, for my example, the 90% C.L. curve is changed only a little. However, curves at lower confidence levels are changed drastically. No hypothesis, including the default hypothesis, has an exclusion level of less than about 50% C.L. This makes sense; if you conduct the experiment once, the data vector you get will typically agree with the default hypothesis better than about 50% of the data vectors that could be generated from that hypothesis.

In pseudocode, the naive algorithm can be represented as

```
Set data vector R to the exact prediction for default hypothesis
For each grid point hypothesis
|  Compute score of R for the grid point hypothesis
|  For N iterations
|  |  Create data vector T, fluctuated from grid point hypothesis
|  |  Compute score of T for the grid point hypothesis
|  |  If score of T is better than score of R, increment a counter n
|  Report result for this grid point: n/N
```

The score may be the likelihood of the data given the hypothesis, the Feldman-Cousins ratio of likelihood to maximum likelihood, or some other measure.

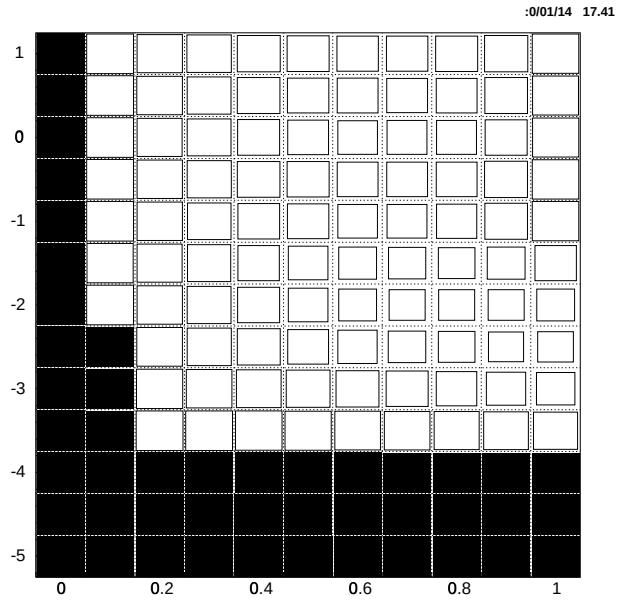

Fig. 3: Plot of exclusion levels using ensemble algorithm.

My ensemble algorithm may be pseudocoded as

```
For i = 1,M
| Create a data vector R_i, fluctuated from default hypothesis
For each grid point hypothesis
|  Compute score of each R_i for the grid point hypothesis
|  For N iterations
|  |  Create data vector T, fluctuated from grid point hypothesis
|  |  Compute score of T for the grid point hypothesis
|  |  For i = 1,M
|  |  |  If score of T is better than score of R_i, increment a counter n
|  Report result for this grid point: n/(N*M)
```

References

[1] G.Feldman and R.Cousins, Phys. Rev. **D57** (1998)3873.

[2] MACRO Collaboration, M. Ambrosio *et al.*, Phys. Lett. **B434** (1998)451.

Discussion after talk of Robert Nolty. Chairman: Peter Igo-Kemenes.

Jacques Bouchez

Don't you think it would be better to publish a median sensitivity curve rather than the mean? The median, that is 50% of the people would find a worse result and 50% a better result, because it avoids this problem of metric dependence.

H. Prosper

Of course, you're quite right in saying that you should use an ensemble, but the result of course depends upon the ensemble that you've used, and depends what you assume to be random and what you assume to be fixed. So in this particular calculation, what did you assume to be fixed and what did you assume to be random?

R. Nolty

I don't know how interesting this is, but I'll do my best to answer that. I assumed that the oscillation parameters were fixed at this value [points on the screen], and I assumed that the absolute normalization was not known, and I treated it as if it were a Gaussian with the mean suggested by our default cross-section calculation and the Bahcall neutrino fluxes, as if it had a Gaussian shape with the errors quoted by the authors of those two models.

J. Linnemann

Were you using only chi-squared to distinguish the theory curves in your angular plot? There are other tests which are more sensitive to the slope, such as the Kolmogorov-Smirnov test.

R. Nolty

In this case I used a complicated prescription that MACRO had come up with, but essentially it was chi-squared. So, maybe other statistics would have done a better job of discriminating these two.